

Exclusion in a priority queue

¹Jan de Gier and ²Caley Finn

Department of Mathematics and Statistics, The University of Melbourne, 3010 VIC, Australia

E-mail: ¹jdger@unimelb.edu.au, ²c.finn3@pgrad.unimelb.edu.au

Abstract. We introduce the prioritising exclusion process, a variation of the totally asymmetric exclusion process with a dynamically varying lattice length. The model acts as a stochastic scheduling mechanism for a priority queueing system, with the varying lattice length representing the length of the queue of customers. We calculate the exact stationary distribution for an unbounded queue by deriving domain wall dynamics from the microscopic transition rules. The structure of the unbounded queue carries over to bounded queues where, although no longer exact, we find the domain wall theory is in very good agreement with simulation results. Within this approximation we calculate average waiting times for queueing customers.

1. Introduction

In this work we introduce a discrete particle hopping and exclusion process motivated by queueing theory: the prioritising exclusion process (PEP). The PEP is an example of a microscopic model of a driven system [1] describing the asymmetric diffusion of hard-core particles along a one-dimensional chain. The empty and occupied sites in the PEP represent low and high priority customers, and the lattice length grows and shrinks as customers arrive and are served. The PEP acts as a stochastic scheduling mechanism for a priority queue, and is closely related to a priority queueing model first introduced by Kleinrock [2], and the subject of more recent work [3].

The hopping and exclusion in the PEP is analogous to that of the totally asymmetric simple exclusion process (TASEP) [4, 5], which is one of the most thoroughly studied and central models of non-equilibrium statistical mechanics [6–9]. The TASEP has been the focus of much mathematical interest due to the fact that it is integrable. Many tools have been applied to or developed for the TASEP. Its exact stationary distribution is known [10, 11] and can be written in matrix product form [9, 12]. Its dynamic properties are studied by means of the Bethe ansatz [8, 13, 14] and for the infinite lattice powerful techniques from random matrix theory are available [15]. Domain wall theory [16] provides a phenomenological explanation of the stationary behaviour of the TASEP. Domain wall theory is amenable to generalisation to more complicated models that may not be integrable, and we will describe it in more detail later.

Recently, several generalisations of the TASEP have been proposed, which allow the lattice length to vary dynamically as is the case in the PEP. These models have been motivated by applications in biological transport [17, 18] and queueing theory [19]. They retain the basic rules of the TASEP: particles occupy sites on a one-dimensional lattice, and hop forward with a given rate to the site immediately ahead, if it is empty. In the biological transport models, particles actively [17] or passively [18] enlarge the domain. In the exclusive queueing process [19], the length of the lattice is defined by the position of the last particle (customer), and the hopping particles represent customers shuffling forwards in the queue.

In the PEP the varying lattice length is defined by the total number of customers at any given time. Aside from its application as a stochastic scheduling mechanism in queueing theory, the PEP is of theoretical interest in statistical physics as it connects systems of different sizes with dynamical rules, contrary to the traditional TASEP where the lattice length is fixed. This fact significantly complicates an exact analysis using techniques such as matrix product states and Bethe ansatz, but we find that the structure of the model makes domain wall theory naturally applicable.

For the PEP, the domain wall dynamics can be derived directly from the microscopic transition rules, similarly to [20]. We find that in the PEP there are two phase transitions. The first one, well known in queueing theory, is between a bounded and unbounded phase where the expected queue length becomes infinite. This transition is independent of the bulk hopping rate and only depends on the arrival and service rates.

The domain wall dynamics yields the exact stationary distribution in the unbounded queue, and gives rise to the second phase transition, which for fixed boundary rates takes place at a critical value of the hopping parameter. At this critical point, the expected jam length of high priority customers near the service end diverges. In other words, beyond this transition, with probability one, low priority customers will no longer be served.

The structure we find in the unbounded phase carries over to the bounded case where, although no longer exact, we see very good agreement with numerical simulations. In the bounded phase a remnant of the second, jamming transition can be observed as a crossover where the jam of high priority customers near the service end delocalises when the hopping parameter reaches a critical value, and the expected jam length becomes comparable to the the queue length.

1.1. The model

In the lattice bulk, the PEP behaves as a TASEP: sites are either occupied by a single particle or empty, and particles hop forwards into empty sites with rate p . At the boundaries the PEP differs from the TASEP. The PEP lattice can be extended on the left by the addition of a filled or empty lattice site, with rates λ_1 and λ_2 respectively. At the other boundary, the rightmost site is removed with rate μ , irrespective of its occupation. These rules, summarised in Figure 1, allow both the lattice length and particle number to vary.

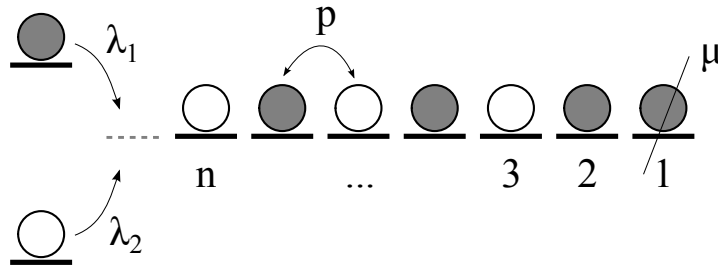


Figure 1. PEP transition rates, filled circles are occupied sites.

We specify a PEP configuration by binary variables τ_i with $\tau_i = 1$ for a filled site and $\tau_i = 0$ for an empty site. Usually we will number sites from right to left and write a length n configuration as

$$\boldsymbol{\tau} = \tau_n \tau_{n-1} \dots \tau_1.$$

1.2. A queueing system

The PEP can be interpreted as a priority queueing system with two classes of customers. The lattice, itself, is the queue of customers, with filled sites representing high priority customers (class 1) and empty sites representing low priority customers (class 2). The

rates λ_1 and λ_2 are the arrival rates of high and low priority customers respectively, and the rate μ , is the rate at which customers are served and leave the queue.

In this interpretation, a particle hopping forward one site corresponds to a high priority customer stepping ahead of the low priority customer immediately in front of them. The stochastic overtaking is the scheduling mechanism in this priority queue, giving high priority customers preferential treatment over low priority. The larger the overtake rate p , the greater the advantage.

The PEP is modelled on a well studied queueing system introduced by Kleinrock [2, 21] and now known as the accumulating priority queue (APQ) [3]. In the APQ, customers have a priority value which accumulates linearly with time. Class 1 customers accumulate priority faster than class 2, thus overtaking them in the service queue. The key difference between the APQ and the PEP is that, for a given sequence of arrivals, overtaking in the APQ is deterministic, but in the PEP the overtakes occur stochastically.

The PEP is also related to a simpler queueing system, the $M/M/1$ queue (see, for example, [22]). The total arrival rate of customers to the PEP is $\lambda = \lambda_1 + \lambda_2$, and the service rate is μ . Both these rates are independent of the internal arrangement of the queue, and the prioritising parameter p . So, if we are interested only in the total length of the queue, we can treat the system as a $M/M/1$ queue with arrival rate λ and service rate μ . The state of an $M/M/1$ queue is characterised simply by the length, n , with probability distribution P_n obeying the master equation

$$\frac{dP_0}{dt} = \mu P_1 - \lambda P_0 \quad (1)$$

$$\frac{dP_n}{dt} = \lambda P_{n-1} + \mu P_{n+1} - (\mu + \lambda) P_n, \quad n > 0. \quad (2)$$

The stationary length distribution of the $M/M/1$ queue (and hence for the PEP) is the solution of $dP_n/dt = 0$, which is

$$P_n = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n, \quad (3)$$

when $\lambda < \mu$, i.e. when the total arrival rate is less than the service rate. In this case the system is described as *stable*, because the queue length does not grow without bound. The expected queue length is finite, given by

$$\langle n \rangle = \frac{\lambda}{\mu - \lambda}. \quad (4)$$

We will call this the *bounded* phase of the PEP.

When $\lambda > \mu$, the system is *unstable* and the expected queue length grows as

$$\langle n \rangle \sim (\lambda - \mu)t. \quad (5)$$

In the late time limit, we can treat the queue as infinite in length. We call this the *unbounded* phase of the PEP.

1.3. Density profiles and waiting times

To define a density profile for the PEP we must specify both the site, i , and the lattice length, n , so that

$$\langle \tau_i \rangle_n = P(\text{queue length is } n, \text{ and site } i \text{ is occupied}).$$

These are one-point functions. We can similarly define higher order correlations

$$\langle \tau_{i_1} \tau_{i_2} \dots \tau_{i_m} \rangle_n, \quad n \geq i_1 > i_2 > \dots > i_m \geq 1. \quad (6)$$

The rate equations for the one-point functions are

$$\frac{d}{dt} \langle \tau_1 \rangle_1 = \lambda_1 P_0 + \mu \langle \tau_2 \rangle_2 - (\lambda + \mu) \langle \tau_1 \rangle_1, \quad (7)$$

$$\frac{d}{dt} \langle \tau_1 \rangle_n = \lambda \langle \tau_1 \rangle_{n-1} + \mu \langle \tau_2 \rangle_{n+1} + p \langle \tau_2 (1 - \tau_1) \rangle_n - (\lambda + \mu) \langle \tau_1 \rangle_n, \quad n > 1, \quad (8)$$

$$\frac{d}{dt} \langle \tau_i \rangle_i = \lambda_1 P_{i-1} + \mu \langle \tau_{i+1} \rangle_{i+1} - p \langle \tau_i (1 - \tau_{i-1}) \rangle_i - (\lambda + \mu) \langle \tau_i \rangle_i, \quad i > 1, \quad (9)$$

$$\begin{aligned} \frac{d}{dt} \langle \tau_i \rangle_n &= \lambda \langle \tau_i \rangle_{n-1} + \mu \langle \tau_{i+1} \rangle_{n+1} + p \langle \tau_{i+1} (1 - \tau_i) \rangle_n \\ &\quad - p \langle \tau_i (1 - \tau_{i-1}) \rangle_n - (\lambda + \mu) \langle \tau_i \rangle_n, \quad i > 1, n > i. \end{aligned} \quad (10)$$

These couple the one-point functions to the two-point correlations, and length n to length $n \pm 1$. The rate equations imply a conserved current of particles across the lattice, but because of the coupling between lengths, some care is required in how this current is defined. We will return to this for the bounded and unbounded phases separately.

Viewing the PEP as a queueing system, we are interested in performance measures, and how these differ for high and low priority customers. The current tells us the rate at which customers pass through the system, and from the density profile we can calculate the average waiting time for customers of each class.

To calculate waiting times, we use Little's result (see Chapter 2.1 of [22]), which states that the average waiting time, \overline{W}_i , is related to the average number of waiting customers, \overline{N}_i , for each class $i = 1, 2$, by

$$\overline{N}_i = \lambda_i \overline{W}_i. \quad (11)$$

The average number of high priority customers can be calculated from the density profile as

$$\overline{N}_1 = \sum_{n=1}^{\infty} \sum_{i=1}^n \langle \tau_i \rangle_n, \quad (12)$$

and the average number of low priority customers is

$$\overline{N}_2 = \langle n \rangle - \overline{N}_1 = \frac{\lambda}{\mu - \lambda} - \overline{N}_1. \quad (13)$$

Here we take the total time from arrival to removal from the system as the waiting time for a customer. Our aim, now, is to compute the density profile for the PEP.

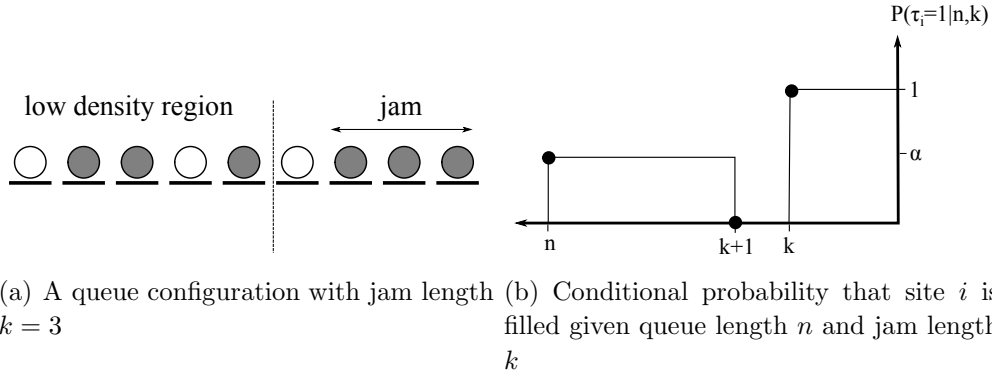


Figure 2.

1.4. Domain wall theory

Domain wall theory [16, 23] reduces the multi-particle dynamics of the TASEP \ddagger to the motion of a single random walker on the lattice. The TASEP boundary conditions (the particle entry and exit rates) create domains of low or high density at the boundaries. These domains extend through the lattice, and where they meet a shock, or domain wall, forms. Domain wall theory models the motion of this shock as a random walk, with the simplifying assumptions that density is constant throughout each domain, and that there is a sharp transition between domains so that the shock can be localised to a single site. Though the stationary solution of the TASEP is known exactly, domain wall theory provides a simple physical explanation of the stationary behaviour [16], and beyond this it allows accurate approximation of some dynamic properties [23]. Domain wall theory can also be applied to more complex models [24, 25] where the exact solution is not known.

The PEP has a natural domain wall structure. As high priority customers overtake and reach the service end, they form a jam (Figure 2(a)): a jam is a section of high priority customers (filled sites) at the service end ahead of any low priority customer (empty site). The jam is characterised by k , the number of consecutive high priority customers. As there are no gaps, there is no overtaking in the jammed region, and the length of the jam reduces only as customers are served. This is similar to the situation in [20], where a TASEP with parallel update and deterministic bulk motion is considered.

Let us assume that the region beyond the jam has uniform density, and that the conditional probability that site i is filled, given the queue length, n , and jam length, k , is (Figure 2(b))

$$P(\tau_i = 1 | n, k) = \begin{cases} 1 & 1 \leq i \leq k \\ 0 & i = k + 1 \\ \alpha & k + 2 \leq i \leq n. \end{cases} \quad (14)$$

\ddagger Domain wall theory applies more generally to the *partially* asymmetric exclusion process.

Then the bulk equation ($n > k + 1, k > 0$) for $P(n, k)$, the probability of a length k jam in a length n queue, is

$$\begin{aligned} \frac{d}{dt}P(n, k) = & \lambda P(n-1, k) + \mu P(n+1, k+1) + \mu(1-\alpha)\alpha^k P(n+1, 0) \\ & + p\alpha P(n, k-1) - (\lambda + \mu + p\alpha)P(n, k). \end{aligned} \quad (15)$$

Let us explain the meaning of each of the terms in equation (15). The term

$$\lambda P(n-1, k)$$

is the entry into the (n, k) configuration from a length n queue due to the arrival of a customer, and the terms

$$\mu P(n+1, k+1) + \mu(1-\alpha)\alpha^k P(n+1, 0),$$

represent the service of a customer. The second term is the transition into the k -jam state from the $(k=0)$ -jam state by the service of a low priority customer who was followed by k consecutive high priority customers. The term

$$p\alpha P(n, k-1),$$

is a $(k-1)$ -jam extending to length k with rate $p\alpha$: there is a high priority customer at site $k+1$ with probability α , which overtakes with rate p the low priority customer at site k (the low priority customer marking the end of the $(k-1)$ -jam). The low priority customer thus moves to position $k+1$ defining the new end of the jam. The loss terms

$$-(\lambda + \mu + p\alpha)P(n, k),$$

are the rate at which the (n, k) configuration is left due to a customer arrival or service, or growth of the jam.

In the next section, we will show that the $n \rightarrow \infty$ limit of (15) follows from the unbounded phase master equation and, together with the equation for $k=0$, leads to the exact solution. In the bounded phase (Section 3), domain wall theory leads to two complementary approximations, which describe the behaviour of the system and allow calculation of waiting times.

2. The unbounded phase

We consider first the unbounded phase of the PEP, where the total arrival rate exceeds the service rate ($\lambda > \mu$). The expected lattice length increases with time, so in the late time limit we can treat it as infinite. On an infinite lattice, the density profile must be given relative to a specified reference frame [17]. For the PEP, there are two reference frames of interest. The *service frame* is fixed at the right hand end of the lattice where customers are served and depart, removing a site from the lattice. The *arrival frame* is fixed at the left hand end of the lattice where customers arrive, adding a filled or empty site to the lattice.

2.1. Domain wall ansatz

To apply domain wall theory to a queue of infinite length, we consider a general but finite section of length m in the service frame, with a jam of length k

$$\tau_m \dots \tau_{k+2} 0 1^k = \boldsymbol{\tau} 0 1^k, \quad (16)$$

where 1^k indicates a string of k 1's. Sites are numbered right to left. The configuration of any finite segment can be written this way, as long as we can take $m \geq k + 1$. We will assume that the conditional probability for a high at site i given a jam of length k is

$$P(\tau_i = 1 | k) = \begin{cases} 1 & 1 \leq i \leq k \\ 0 & i = k + 1 \\ \alpha & k + 2 \leq i, \end{cases} \quad (17)$$

which is (14) in the $n \rightarrow \infty$ limit. Then the probability of the finite segment (16) is

$$\begin{aligned} P(\boldsymbol{\tau} 0 1^k) &= \sum_{\tau_\infty, \dots, \tau_{m+1}=0,1} P(\dots \tau_{m+1} \tau_m \dots \tau_{k+2} 0 1^k) \\ &= \alpha^h (1 - \alpha)^l P_{\text{jam}}(k), \end{aligned} \quad (18)$$

where h is the number of highs in the configuration beyond the jam up to position m , and l is the number of lows, that is

$$h = \sum_{i=k+2}^m \tau_i, \quad l = m - h - k - 1, \quad (19)$$

and $P_{\text{jam}}(k)$ is the probability of a length k jam. The jam probabilities are normalised such that

$$\sum_{k=0}^{\infty} P_{\text{jam}}(k) = 1. \quad (20)$$

Equation (18) is an ansatz for the stationary probabilities. We now show that α and $P_{\text{jam}}(k)$ may be determined such that all rate equations are satisfied.

2.2. The service frame

We use the notation $\boldsymbol{\tau}|_{(i,i-1)}$ to indicate the exchange of customers in places i and $i - 1$. That is, for $\boldsymbol{\tau} = \tau_r \dots \tau_1$

$$\boldsymbol{\tau}|_{(i,i-1)} = \tau_r \dots \tau_{i+1} \tau_{i-1} \tau_i \tau_{i-2} \dots \tau_1,$$

and

$$0 \boldsymbol{\tau}|_{(r+1,r)} = \tau_r 0 \tau_{r-1} \dots \tau_1.$$

The stationary rate equation for the k -jam configuration (16) with $k \geq 1$ is

$$\begin{aligned} 0 &= \frac{d}{dt} P(\boldsymbol{\tau} 0 1^k) \\ &= \mu P(\boldsymbol{\tau} 0 1^{k+1}) + \mu P(\boldsymbol{\tau} 0 1^k 0) \end{aligned}$$

$$\begin{aligned}
& + p\tau_m P\left(0\tau 01^k|_{(m+1,m)}\right) + \sum_{i=k+2}^m p(1-\tau_i)\tau_{i-1} P\left(\tau 01^k|_{(i,i-1)}\right) + pP\left(\tau 101^{k-1}\right) \\
& - \mu P(\tau 01^k) - \sum_{i=k+2}^m p\tau_i(1-\tau_{i-1})P(\tau 01^k) - p(1-\tau_m)P(1\tau 01^k).
\end{aligned} \tag{21}$$

Let us again explain the various terms. The terms

$$\mu P\left(\tau 01^{k+1}\right) + \mu P\left(\tau 01^k 0\right),$$

give the rate of arrival to the k -jam configuration after, respectively, a high or low priority customer is served. Then there are the hopping terms. A high in m th place in $\tau 01^k$ can arrive from place $m+1$:

$$p\tau_m P\left(0\tau 01^k|_{(m+1,m)}\right).$$

Overtaking within the low density region behind the jam is given by

$$\sum_{i=k+2}^m p(1-\tau_i)\tau_{i-1} P\left(\tau 01^k|_{(i,i-1)}\right),$$

and a $(k-1)$ -jam extends to a k -jam when a high hops onto the end:

$$pP\left(\tau 101^{k-1}\right).$$

The loss term

$$-\mu P(\tau 01^k),$$

is the reduction of the jam as a customer is served, and

$$- \sum_{i=k+2}^m p\tau_i(1-\tau_{i-1})P(\tau 01^k)$$

are overtakings within the m places of $\tau 01^k$. The final loss term

$$-p(1-\tau_m)P(1\tau 01^k)$$

arises if the configuration has a low in m th place, which can be overtaken by a high from place $m+1$.

Substituting the ansatz (18), the terms representing overtaking within the m sites of $\tau 01^k$ combine and telescope to

$$\begin{aligned}
& p \left(\sum_{i=k+2}^m (1-\tau_i)\tau_{i-1} - \sum_{i=k+2}^m \tau_i(1-\tau_{i-1}) \right) \alpha^h(1-\alpha)^l P_{\text{jam}}(k) \\
& = p(\tau_{k+1} - \tau_m) \alpha^h(1-\alpha)^l P_{\text{jam}}(k) \\
& = -p\tau_m \alpha^h(1-\alpha)^l P_{\text{jam}}(k);
\end{aligned}$$

recall that $\tau_{k+1} = 0$.

The factor $\alpha^h(1-\alpha)^l$ is common to all terms in the rate equation. Cancelling, and simplifying leaves

$$0 = p\alpha P_{\text{jam}}(k-1) + \mu P_{\text{jam}}(k+1) + \mu(1-\alpha)\alpha^k P_{\text{jam}}(0) - (\mu + p\alpha)P_{\text{jam}}(k). \tag{22}$$

This agrees with the $n \rightarrow \infty$ limit of (15), but we have derived it from the full PEP rate equations. Were it not for the $P_{\text{jam}}(0)$ term, this equation for the position of the jam would have the same form as the domain wall theory for the TASEP [23].

The $k = 0$ case differs only slightly. In this case the rate equation is

$$\begin{aligned} 0 &= \frac{d}{dt} P(\tau 0) \\ &= \mu P(\tau 01) + \mu P(\tau 00) + p\tau_m P(0\tau 0|_{(m+1,m)}) + \sum_{i=2}^m p(1-\tau_i)\tau_{i-1} P(\tau 0|_{(i,i-1)}) \\ &\quad - \mu P(\tau 0) - \sum_{i=2}^m p\tau_i(1-\tau_{i-1}) P(\tau 0) - p(1-\tau_m) P(1\tau 0) \end{aligned} \quad (23)$$

Substituting the ansatz (18), this reduces to

$$0 = \mu P_{\text{jam}}(1) - (\mu\alpha + p\alpha) P_{\text{jam}}(0), \quad (24)$$

and rearranging gives

$$P_{\text{jam}}(1) = \frac{p\alpha}{\mu} P_{\text{jam}}(0) + \alpha P_{\text{jam}}(0). \quad (25)$$

With this as the base case, we use (22) to show by induction that

$$P_{\text{jam}}(k) = \frac{p\alpha}{\mu} P_{\text{jam}}(k-1) + \alpha^k P_{\text{jam}}(0), \quad k \geq 1. \quad (26)$$

This recurrence for $P_{\text{jam}}(k)$ has solution

$$\begin{aligned} P_{\text{jam}}(k) &= \sum_{i=0}^k \left(\frac{p\alpha}{\mu} \right)^{k-i} \alpha^i P_{\text{jam}}(0) \\ &= \frac{p \left(\frac{p\alpha}{\mu} \right)^k - \mu \alpha^k}{p - \mu} P_{\text{jam}}(0). \end{aligned} \quad (27)$$

The normalisation condition (20) fixes

$$P_{\text{jam}}(0) = (1 - \alpha) \left(1 - \frac{p\alpha}{\mu} \right), \quad (28)$$

subject to the constraint

$$p\alpha < \mu. \quad (29)$$

The domain wall picture makes the meaning of this constraint clear. The jam of high priority customers grows with rate $p\alpha$ and is reduced with rate μ . If $p\alpha > \mu$ the jam grows with rate $p\alpha - \mu > 0$, that is

$$\langle k \rangle \sim (p\alpha - \mu)t, \quad (30)$$

and as $t \rightarrow \infty$ the expected length of the jam becomes infinite. In contrast, when (29) is satisfied, the service rate is fast enough to prevent a backlog of high priority customers, and the expected jam length is finite and given by

$$\langle k \rangle = \sum_{k=1}^{\infty} k P_{\text{jam}}(k) = \frac{\alpha}{1 - \alpha} + \frac{p\alpha}{\mu - p\alpha}. \quad (31)$$

2.3. The arrival frame

To determine α , we examine the PEP in the arrival frame. Sites are now numbered left to right, and the interchange operation is defined as

$$\boldsymbol{\tau}|_{(i,i+1)} = \tau_1 \dots \tau_{i-1} \tau_{i+1} \tau_i \tau_{i+2} \dots \tau_r,$$

for $\boldsymbol{\tau} = \tau_1 \dots \tau_r$.

We will assume that the jam is always far from the arrival end. In the late time limit this is guaranteed if

$$p\alpha < \lambda. \quad (32)$$

This condition is clearly met when (29) is satisfied, but we will show that α can be determined consistently with this requirement. Then in the arrival frame, the ansatz (18) implies

$$P(\tau_i = 1) = \alpha, \quad P(\tau_i = 0) = 1 - \alpha,$$

for any site. A configuration on the first m sites,

$$\boldsymbol{\tau} = \tau_1 \dots \tau_m,$$

occurs with probability

$$P(\boldsymbol{\tau}) = \alpha^h (1 - \alpha)^l, \quad (33)$$

where

$$h = \sum_{i=1}^m \tau_i, \quad l = m - h.$$

The stationary rate equation for this configuration is

$$\begin{aligned} 0 &= \frac{d}{dt} P(\boldsymbol{\tau}) \\ &= \lambda_1 \tau_1 P(\tau_2 \dots \tau_m) + \lambda_2 (1 - \tau_1) P(\tau_2 \dots \tau_m) \\ &\quad + \sum_{i=1}^{m-1} p(1 - \tau_i) \tau_{i+1} P(\boldsymbol{\tau}|_{(i,i+1)}) + p(1 - \tau_m) P(\boldsymbol{\tau} 1_{(m,m+1)}) \\ &\quad - \lambda P(\boldsymbol{\tau}) - \sum_{i=1}^{m-1} p \tau_i (1 - \tau_{i+1}) P(\boldsymbol{\tau}) - p \tau_m P(\boldsymbol{\tau} 0). \end{aligned} \quad (34)$$

Substituting (33), the summed hopping terms again combine and telescope, and the factors of α and $1 - \alpha$ common to all terms can be cancelled. This leaves

$$0 = -\lambda\alpha(1 - \alpha) + p\alpha^2(1 - \alpha) - p\tau_1\alpha(1 - \alpha) + \tau_1\lambda_1(1 - \alpha) + (1 - \tau_1)\lambda_2\alpha,$$

which for both $\tau_1 = 0$ and $\tau_1 = 1$ reduces to

$$p\alpha^2 - (p + \lambda)\alpha + \lambda_1 = 0.$$

The two solutions are

$$\alpha_{\pm} = \frac{p + \lambda \pm \sqrt{(p - \lambda)^2 + 4p\lambda_2}}{2p}, \quad (35)$$

with $0 < \alpha_- < 1$ and $\alpha_+ > 1$ for $p, \lambda_1, \lambda_2 > 0$. As α is a density value we must take $\alpha = \alpha_-$. Substituting this value for α , it is seen that $p\alpha < \lambda$ (the constraint (32)) is satisfied for all physical parameter values. Thus, starting from the assumption that the jam never reaches the arrival end, we arrive at a consistent solution.

2.4. Density profile, conserved currents, and service rates

To obtain the service frame rate equations, we take the $n \rightarrow \infty$ limit of (7–10). In this limit, the boundary cases can be neglected, and the bulk equations can be written

$$\frac{d}{dt} \langle \tau_i \rangle_\infty = J_\infty^{(i+1)} - J_\infty^{(i)}, \quad i \geq 1, \quad (36)$$

where

$$J_\infty^{(1)} = \mu \langle \tau_1 \rangle_\infty \quad (37)$$

$$J_\infty^{(i)} = \mu \langle \tau_i \rangle_\infty + p \langle \tau_i (1 - \tau_{i-1}) \rangle_\infty, \quad i \geq 2. \quad (38)$$

In the stationary state the time derivatives are zero so (36) defines a conserved current

$$J_\infty = J_\infty^{(1)} = J_\infty^{(2)} = \dots \quad (39)$$

The bulk current, $J_\infty^{(i)}$, has the usual TASEP hopping term, $p \langle \tau_i (1 - \tau_{i-1}) \rangle_\infty$. The additional term $\mu \langle \tau_i \rangle_\infty$ arises due to the choice of reference frame.

The one-point functions are computed from the domain wall solution as

$$\begin{aligned} \langle \tau_i \rangle_\infty &= \alpha \sum_{k=0}^{i-2} P_{\text{jam}}(k) + \sum_{k=i}^{\infty} P_{\text{jam}}(k) \\ &= \alpha + (1 - \alpha) \left(\frac{p\alpha}{\mu} \right)^i, \end{aligned} \quad (40)$$

and the two-point correlation $\langle \tau_{i+1} (1 - \tau_i) \rangle_\infty$ is

$$\begin{aligned} \langle \tau_{i+1} (1 - \tau_i) \rangle_\infty &= \alpha (1 - \alpha) \sum_{k=0}^{i-2} P_{\text{jam}}(k) + \alpha P_{\text{jam}}(i-1) \\ &= \alpha (1 - \alpha) \left(1 - \left(\frac{p\alpha}{\mu} \right)^i \right). \end{aligned} \quad (41)$$

The resulting service frame current is

$$J_\infty = p\alpha(1 - \alpha) + \mu\alpha, \quad (42)$$

and is the rate at which particles exit the system.

In terms of the queueing model, the current is the average rate at which high priority customers leave the queue. As the total rate at which customers leave the system is μ , § low priority customers leave the queue at rate

$$\mu - J_\infty = (\mu - p\alpha)(1 - \alpha). \quad (43)$$

§ If the queue was ever empty, the rate at which customers leave would be less than the service rate, but in the unbounded phase this is not a concern.

The constraint $p\alpha < \mu$ (equation (29)) ensures that this rate is greater than zero and low priority customers always receive a share of the service. For $p\alpha > \mu$, the jam of high priority customers at the service end becomes unbounded, and low priority customers can no longer reach the front of the queue to be served. Thus the low priority current has a second order phase transition at $p\alpha = \mu$.

The phase transition subdivides the unbounded phase of the PEP. By fixing values for $\lambda = \lambda_1 + \lambda_2$ and μ , we can plot illustrative two dimensional phase diagrams with λ_1 and p as the axes. Using (35) for $\alpha = \alpha_-$, the curve where $p\alpha = \mu$ is given by

$$\lambda_1^{(\infty)}(p) = \begin{cases} \lambda & p < \mu \\ \mu \left(1 + \frac{\lambda - \mu}{p} \right) & p \geq \mu, \end{cases} \quad (44)$$

and $p\alpha < \mu$ for $\lambda_1 < \lambda_1^{(\infty)}(p)$. Figure 3 shows the phase diagram for $\lambda = 1.5, \mu = 1$. The

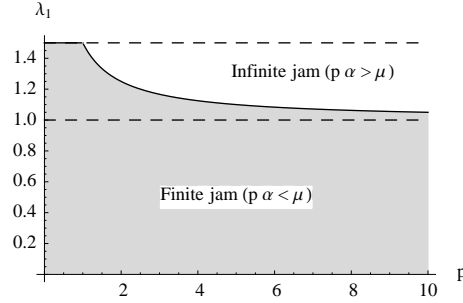


Figure 3. Subdivisions of the unbounded phase with $\lambda = 1.5, \mu = 1$.

function $\lambda_1^{(\infty)}(p)$ is decreasing in p and $\lim_{p \rightarrow \infty} \lambda_1(p) = \mu$. Therefore, the transition into the ‘infinite jam’ phase occurs only if $\lambda_1 > \mu$. This is marked by the lower dashed line. The upper dashed line marks the $\lambda_1 = \lambda, \lambda_2 = 0$ boundary.

3. The bounded phase

In the queueing theory interpretation, it is the bounded phase of the PEP (with $\lambda < \mu$) that is of greatest interest. In this phase the queue lengths and waiting times remain finite, and we can compare how the waiting time varies with the overtake rate, p , for high and low priority class customers.

The fluctuating lattice length proves a challenge in applying domain wall theory directly to the bounded phase. We will take two approaches, each leading to an approximate solution revealing different aspects of the system. The first method tells us about the shape and length dependence of the density profiles, while the second method allows us to calculate customer waiting times.

3.1. Domain wall ansatz

To apply the domain wall ansatz directly in the bounded phase, we consider a general length n configuration with a length k jam,

$$\tau_n \dots \tau_{k+2} 0 1^k = \boldsymbol{\tau} 0 1^k. \quad (45)$$

The stationary rate equation for $k > 0$, $n > k + 1$ is

$$\begin{aligned} 0 &= \frac{d}{dt} P(\boldsymbol{\tau} 0 1^k) \\ &= \lambda_1 \tau_n P(\tau_{n-1} \dots \tau_{k+2} 0 1^k) + \lambda_2 (1 - \tau_n) P(\tau_{n-1} \dots \tau_{k+2} 0 1^k) \\ &\quad + \sum_{i=k+2}^n p (1 - \tau_i) \tau_{i-1} P(\boldsymbol{\tau} 0 1^k |_{(i,i-1)}) + p P(\boldsymbol{\tau} 1 0 1^{k-1}) + \mu P(\boldsymbol{\tau} 0 1^{k+1}) + \mu P(\boldsymbol{\tau} 0 1^k 0) \\ &\quad - (\lambda + \mu) P(\boldsymbol{\tau} 0 1^k) - \sum_{i=k+2}^n p \tau_i (1 - \tau_{i-1}) P(\boldsymbol{\tau} 0 1^k), \end{aligned} \quad (46)$$

and for $k = 0$, $n > 1$

$$\begin{aligned} 0 &= \frac{d}{dt} P(\boldsymbol{\tau} 0) \\ &= \lambda_1 \tau_n P(\tau_{n-1} \dots \tau_2 0) + \lambda_2 (1 - \tau_n) P(\tau_{n-1} \dots \tau_2 0) \\ &\quad + \sum_{i=2}^n p (1 - \tau_i) \tau_{i-1} P(\boldsymbol{\tau} 0 |_{(i,i-1)}) + \mu P(\boldsymbol{\tau} 0 1) + \mu P(\boldsymbol{\tau} 0 0) \\ &\quad - (\lambda + \mu) P(\boldsymbol{\tau} 0) - \sum_{i=2}^n p \tau_i (1 - \tau_{i-1}) P(\boldsymbol{\tau} 0). \end{aligned} \quad (47)$$

These are “bulk” equations, valid when the jam is away from the arrival end of the queue and are of the form discussed in Section 1.4. To see this, define

$$P(n, k) = \sum_{\tau_n, \dots, \tau_{k+2}=0,1} P(\tau_n \dots \tau_{k+2} 0 1^k), \quad (48)$$

which is the probability of a length k jam in a length n queue. For $k > 0$, $n > k + 1$, summing (46) and applying the domain wall ansatz (14) gives

$$\begin{aligned} 0 &= \frac{d}{dt} P(n, k) = \lambda P(n-1, k) + \mu P(n+1, k+1) + \mu(1-\alpha)\alpha^k P(n+1, 0) \\ &\quad + p\alpha P(n, k-1) - (\lambda + \mu + p\alpha) P(n, k). \end{aligned} \quad (49)$$

which is exactly (15). And for $k = 0$, $n > 1$, summing (47) gives

$$\begin{aligned} 0 &= \frac{d}{dt} P(n, 0) \\ &= \lambda P(n-1, 0) + \mu P(n+1, 1) + \mu(1-\alpha) P(n+1, 0) - (\lambda + \mu + p\alpha) P(n, 0). \end{aligned} \quad (50)$$

The simple domain wall picture breaks down when the jam reaches the arrival end, i.e. for configurations $0 1^{n-1}$ or 1^n . Our strategy is to find a solution of the bulk equations, without requiring it to satisfy these boundary equations. We can hope that this will give an approximation to the true solution. What we will show is that, within the range of validity, the approximation is very good.

3.1.1. Length assumption. To solve the bulk equations we assume the length dependence factorises as

$$P(n, k) = P_n P_{\text{jam}}^*(k), \quad (51)$$

where P_n is the length distribution (3). Then equation (49), for $k > 0$, becomes

$$0 = p\alpha P_{\text{jam}}^*(k-1) + \lambda P_{\text{jam}}^*(k+1) + \lambda(1-\alpha)\alpha^k P_{\text{jam}}^*(0) - (\lambda + p\alpha)P_{\text{jam}}^*(k), \quad (52)$$

and equation (50), for $k = 0$, gives

$$0 = \lambda P_{\text{jam}}^*(1) - (\lambda\alpha + p\alpha)P_{\text{jam}}^*(0). \quad (53)$$

These have the same form as the unbounded queue domain wall equations, (22), (24), but with λ in place of μ . Therefore they are solved by

$$\begin{aligned} P_{\text{jam}}^*(k) &= \sum_{i=0}^k \left(\frac{p\alpha}{\lambda}\right)^{k-i} \alpha^i P_{\text{jam}}^*(0) \\ &= \frac{p\left(\frac{p\alpha}{\lambda}\right)^k - \lambda\alpha^k}{p - \lambda} P_{\text{jam}}^*(0). \end{aligned} \quad (54)$$

The normalisation of $P_{\text{jam}}^*(k)$ must be independent of n . For the solution to be valid for $n \rightarrow \infty$ (as there is no cap on queue length) we must require

$$\sum_{k=0}^{\infty} P_{\text{jam}}^*(k) = 1, \quad (55)$$

fixing

$$P_{\text{jam}}^*(0) = (1 - \alpha) \left(1 - \frac{p\alpha}{\lambda}\right), \quad (56)$$

subject to the constraint

$$p\alpha < \lambda. \quad (57)$$

The total probability at each length, n , must sum to the length distribution (3), that is

$$\sum_{k=0}^{n-1} P(n, k) + P(1^n) = P_n. \quad (58)$$

As only $P(1^n)$ is undertermined, we must have that

$$\begin{aligned} P(1^n) &= P_n - \sum_{k=0}^{n-1} P(n, k) \\ &= P_n \sum_{k=n}^{\infty} P_{\text{jam}}^*(k) \\ &= P_n \frac{p(1 - \alpha) \left(\frac{p\alpha}{\lambda}\right)^n - \lambda \left(1 - \frac{p\alpha}{\lambda}\right) \alpha^n}{p - \lambda}. \end{aligned} \quad (59)$$

This is analogous to the unbounded queue. There the probability that the first n sites from the service end are filled is $\sum_{k=n}^{\infty} P_{\text{jam}}(k)$.

To determine α we return to the general k -jam equation (46)||, and apply the domain wall ansatz (14) and the length assumption (51), leaving

$$0 = \frac{\lambda_1 \mu}{\lambda} \tau_n (1 - \alpha) P_{\text{jam}}^*(k) + \frac{\lambda_2 \mu}{\lambda} (1 - \tau_n) \alpha P_{\text{jam}}^*(k) + p \alpha^2 (1 - \alpha) P_{\text{jam}}^*(k - 1) \\ + \lambda \alpha (1 - \alpha) P_{\text{jam}}^*(k + 1) + \lambda (1 - \alpha)^2 \alpha^{k+1} P_{\text{jam}}^*(0) \\ - (\lambda + \mu + p \tau_n) \alpha (1 - \alpha) P_{\text{jam}}^*(k). \quad (60)$$

Multiplying (52) by $\alpha(1 - \alpha)$ and subtracting from (60), we then consider $\tau_n = 0$ and $\tau_n = 1$ separately. Both cases reduce to

$$0 = p \alpha^2 - (p + \mu) \alpha + \frac{\lambda_1 \mu}{\lambda}, \quad (61)$$

with solutions

$$\alpha_{\pm} = \frac{p + \mu \pm \sqrt{(p + \mu)^2 - 4p \frac{\lambda_1 \mu}{\lambda}}}{2p}. \quad (62)$$

Using the inequalities

$$\lambda_1 < \frac{\lambda_1 \mu}{\lambda} < \mu, \quad (63)$$

(the first inequality holds as the queue is bounded) we see that $0 < \alpha_- < 1$, and $\alpha_+ > 1$ when $p, \lambda_1, \lambda_2, \mu > 0$. Again, we must take $\alpha = \alpha_-$ to have a proper density value.

3.1.2. Density profile. To summarise, we have solved the bulk equations in the domain wall approximation, giving the solution in the form (51), which with (3) and (54) results in

$$P(n, k) = \left(1 - \frac{\lambda}{\mu}\right) \left(1 - \frac{p\alpha}{\lambda}\right) \frac{1 - \alpha}{p - \lambda} \left(\frac{\lambda}{\mu}\right)^n \left(p \left(\frac{p\alpha}{\lambda}\right)^k - \lambda \alpha^k\right). \quad (64)$$

In general this solution does not satisfy the boundary equations for $k = n, n - 1$, i.e. the cases we neglected were where the jam extends to the length of the queue. The jam grows with rate $p\alpha$, so the constraint $p\alpha < \lambda$ (equation (57) arising from the normalisation condition) requires that the queue length grows faster on average than the jam. Since we have totally neglected the boundary equations, we expect our approximation to be best when $p\alpha \ll \lambda$.

The density at site i in a length n queue, computed from this solution is

$$\langle \tau_i \rangle_n = \alpha \sum_{k=0}^{i-2} P(n, k) + \sum_{k=0}^{n-1} P(n, k) + P(1^n) \\ = \alpha \sum_{k=0}^{i-2} P(n, k) + \sum_{k=0}^{\infty} P(n, k) \\ = P_n \left(\alpha + (1 - \alpha) \left(\frac{p\alpha}{\lambda}\right)^i \right). \quad (65)$$

Figure 4 shows length dependent density profiles for $p\alpha < \lambda$ (Figure 4(a)) and $p\alpha > \lambda$ (Figure 4(b)). Triangle markers show Monte Carlo simulation results, while the solid

|| This is for $k > 0$, but (47) for $k = 0$ gives the same result.

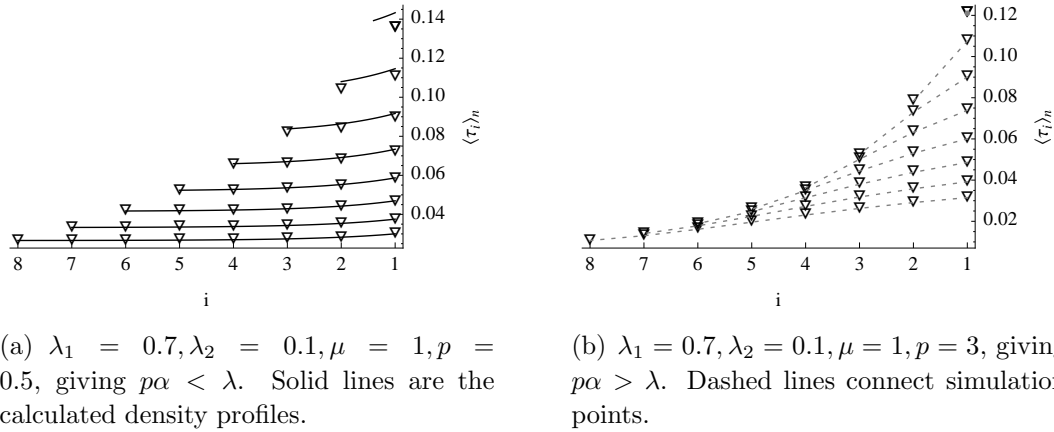


Figure 4. Density profiles for lengths 1 – 8 showing differing length dependence. Triangle markers for simulation results.

curves in Figure 4(a) are the density profiles $\langle \tau_i \rangle_n$ calculated from (65). The simulation points in Figure 4(b) have been connected by dashed lines to aid the eye.

In Figure 4(a), we see good agreement between simulated and calculated profiles, the discrepancy between the two decreasing as n increases. But what is most striking is the difference in length dependence between Figures 4(a) and 4(b), with the length factorisation assumption (51) clearly not holding in 4(b). The almost linear profiles in 4(b) are reminiscent of the behaviour of the TASEP on the coexistence line. In that phase, the domain wall occurs at all positions with equal probability, resulting, on average, in a linear profile [11]. In the bounded PEP the evidence suggests that the constraint $p\alpha < \lambda$ marks a crossover between the localised jam, for $p\alpha < \lambda$ and the delocalised jam for $p\alpha > \lambda$.

3.2. Aggregate density profile and current

In this section we take a different approach, working with the rate equations for the one-point functions. We start by defining a conserved current for the bounded phase. To do so, we sum the density at each position over all lengths, thus aggregating the effect of the length fluctuations. Define the summed one-point functions

$$\overline{\langle \tau_i \rangle} = \sum_{n=i}^{\infty} \langle \tau_i \rangle_n. \quad (66)$$

Note that as $\langle \tau_i \rangle_n \leq P_n$,

$$\overline{\langle \tau_i \rangle} \leq \sum_{n=i}^{\infty} P_n = \left(\frac{\lambda}{\mu} \right)^i, \quad (67)$$

so the sum is bounded, and it converges as $\sum_{n=i}^M \langle \tau_i \rangle_n$ is monotone increasing in M . Thus $\overline{\langle \tau_i \rangle}$ is well defined. Higher order summed correlations are defined similarly, e.g.

$$\overline{\langle \tau_{i+1}(1 - \tau_i) \rangle} = \sum_{n=i}^{\infty} \langle \tau_{i+1}(1 - \tau_i) \rangle_n. \quad (68)$$

Summing the rate equations (7) – (10) gives

$$\frac{d}{dt}\langle\tau_1\rangle = \lambda_1 P_0 + \mu\langle\tau_2\rangle + p\langle\tau_2(1-\tau_1)\rangle - \mu\langle\tau_1\rangle, \quad (69)$$

$$\frac{d}{dt}\langle\tau_i\rangle = \lambda_1 P_{i-1} + \mu\langle\tau_{i+1}\rangle + p\langle\tau_{i+1}(1-\tau_i)\rangle - p\langle\tau_i(1-\tau_{i-1})\rangle - \mu\langle\tau_i\rangle, \quad i > 1. \quad (70)$$

These can be written as a conservation equation for three currents

$$\frac{d}{dt}\langle\tau_i\rangle = J_{\text{ext}}^{(i)} + \overline{J}^{(i+1,i)} - \overline{J}^{(i,i-1)}, \quad i \geq 1, \quad (71)$$

where

$$J_{\text{ext}}^{(i)} = \lambda_1 P_{i-1} \quad (72)$$

$$\overline{J}^{(1,0)} = \mu\langle\tau_1\rangle \quad (73)$$

$$\overline{J}^{(i,i-1)} = \mu\langle\tau_i\rangle + p\langle\tau_i(1-\tau_{i-1})\rangle, \quad i \geq 2. \quad (74)$$

$\overline{J}^{(i,i-1)}$ is the site-to-site current with a hopping term and frame current term. But customers can also step directly into place at the end of the queue, which gives the external current $J_{\text{ext}}^{(i)}$.

In the stationary distribution the time derivatives are zero, and so (71) gives a recurrence for the site-to-site current

$$\overline{J}^{(i+1,i)} = \overline{J}^{(i,i-1)} - J_{\text{ext}}^{(i)}, \quad (75)$$

which reduces to

$$\overline{J}^{(i+1,i)} = \overline{J}^{(1,0)} - \lambda_1 \sum_{n=0}^{i-1} P_n. \quad (76)$$

We have $\lim_{i \rightarrow \infty} \overline{J}^{(i+1,i)} = 0$, as both terms on the right hand side of (74) go to zero. Therefore (76) implies that

$$\overline{J}^{(1,0)} = \lambda_1 \sum_{n=0}^{\infty} P_n = \lambda_1, \quad (77)$$

and

$$\overline{J}^{(i,i-1)} = \lambda_1 \left(1 - \sum_{n=0}^{i-2} P_n\right) = \lambda_1 \left(\frac{\lambda}{\mu}\right)^{i-1}, \quad i \geq 1. \quad (78)$$

Now (73) tells us that

$$\langle\tau_1\rangle = \frac{\lambda_1}{\mu}. \quad (79)$$

This result is exact – it is the probability that the lattice is at least length one with a particle in site 1.

We could have derived this result directly from the notion of the PEP as a queue: high priority customers leave the system at an average rate $\mu\langle\tau_1\rangle$. The rate at which high priority customers leave the system cannot be higher than λ_1 , the rate they arrive. But as the service capacity exceeds the total arrival rate ($\lambda = \lambda_1 + \lambda_2$) there is no

bottleneck at the server, so high priority customers¶ leave at the rate they arrive, that is $\mu\overline{\langle\tau_1\rangle} = \lambda_1$.

The presence of two-point correlations prevent us from calculating exact densities for $i = 2, 3$, etc. The mean field method [10, 17] is the standard way to deal with this, assuming the correlation between neighbouring sites is small so that the two point correlations can be approximated as products of the one-point functions. But for the PEP, we can exploit the similarity to the unbounded system, for which we have an exact solution.

By the definition (66), $\overline{\langle\tau_i\rangle}$ is the probability

$$\overline{\langle\tau_i\rangle} = P(\tau_i = 1, \text{length } n \geq i). \quad (80)$$

We can instead work with the conditional probability

$$\langle\tau_i|n \geq i\rangle = P(\tau_i = 1|\text{length } n \geq i); \quad (81)$$

the two are related by

$$\overline{\langle\tau_i\rangle} = P(\text{length } n \geq i)\langle\tau_i|n \geq i\rangle = \left(\frac{\lambda}{\mu}\right)^i \langle\tau_i|n \geq i\rangle. \quad (82)$$

Similarly, define $\langle\tau_i(1 - \tau_{i-1})|n \geq i\rangle$ through

$$\overline{\langle\tau_i(1 - \tau_{i-1})\rangle} = \left(\frac{\lambda}{\mu}\right)^i \langle\tau_i(1 - \tau_i)|n \geq i\rangle. \quad (83)$$

Substituting into (73), (74) gives

$$\begin{aligned} \frac{\lambda_1\mu}{\lambda} &= \mu\langle\tau_1|n \geq 1\rangle \\ &= \mu\langle\tau_i|n \geq i\rangle + p\langle\tau_i(1 - \tau_{i-1})|n \geq i\rangle, \quad i \geq 2. \end{aligned} \quad (84)$$

These have the same form as the current equations for the unbounded queue, (38), with an effective current

$$\tilde{J} = \frac{\lambda_1\mu}{\lambda}, \quad (85)$$

so are solved by the unbounded queue one- and two-point functions (40), (41). That is

$$\langle\tau_i|n \geq i\rangle = \alpha + (1 - \alpha) \left(\frac{p\alpha}{\mu}\right)^i, \quad i \geq 1, \quad (86)$$

and

$$\langle\tau_i(1 - \tau_{i-1})|n \geq i\rangle = \alpha(1 - \alpha) \left(1 - \left(\frac{p\alpha}{\mu}\right)^{i-1}\right), \quad i \geq 2. \quad (87)$$

To determine α , we substitute (86), (87) into (84), and take $i \rightarrow \infty$ (assuming $p\alpha < \mu$). This gives back the quadratic for α (61), so we again must take $\alpha = \alpha_-$, given by (62). Note that $p\alpha < \mu$ if $p, \mu, \lambda_2 > 0$, so the density profile is always exponentially decaying.

Figure 5 shows the aggregated density profiles, $\overline{\langle\tau_i\rangle}$ computed from (82), (86), plotted against simulation results for several different parameter values. The shape

¶ A corresponding argument applies to low priority customers.

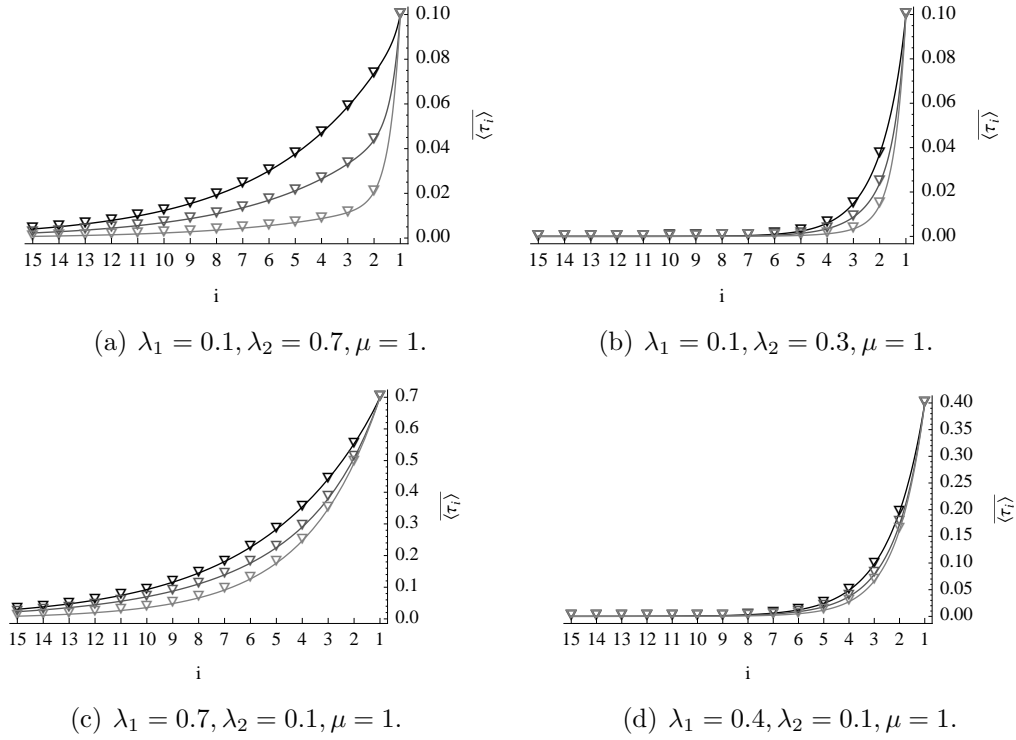


Figure 5. Summed density profiles. Triangle markers for simulation results, plotted against calculated profile $\langle \tau_i \rangle$. For $p = 0.1$ (black), $p = 1$ (mid-gray), $p = 5$ (light gray).

of the profiles, and dependence on the overtake rate, p , can be understood in terms of the domain wall model. For $i = 1$, where the exact value of $\langle \tau_1 \rangle$ is known, and for large i , where the length dependence in the form of the $(\lambda/\mu)^i$ factor dominates, there is excellent agreement between the simulated and calculated density profiles. At intermediate values of i the dependence on p is seen. In the calculated density profile, this dependence enters through the $(p\alpha/\mu)^i$ term, which can be understood as describing a jam growing with rate $p\alpha$ and being reduced with rate μ . It is here that there is the greatest discrepancy between calculated and simulated profiles, although the agreement is still very good.

Note we could also compare the aggregated profiles with $\langle \tau_i \rangle_n$ (65) summed over n . However, even at position 1 the summed $\langle \tau_1 \rangle_n$ does not agree with $\langle \tau_1 \rangle$ (79) for which the exact result is known. The direct application of the domain wall ansatz in Section 3.1 gave an indication of the length dependence in the system, but the approach in this section, following from the current conservation equation, is in much better agreement numerically with simulation results.

3.3. Waiting times

We can use the aggregated one-point functions to compute the average number of customers in the queue, and in turn the waiting times for both classes of customers.

Switching the order of the sums in (12), we can write \overline{N}_1 , the average number of high priority customers, as

$$\overline{N}_1 = \sum_{i=1}^{\infty} \sum_{n=i}^{\infty} \langle \tau_i \rangle_n = \sum_{i=1}^{\infty} \overline{\langle \tau_i \rangle}. \quad (88)$$

Substituting (82), (86) into (88), Little's result (11) gives the average high priority waiting time,

$$\overline{W}_1 = \frac{1}{\lambda_1} \overline{N}_1 = \frac{1}{\lambda_1} \left(\alpha \frac{\lambda}{\mu - \lambda} + (1 - \alpha) \frac{p\alpha\lambda}{\mu^2 - p\alpha\lambda} \right). \quad (89)$$

With (13), the average low priority waiting time is

$$\overline{W}_2 = \frac{1}{\lambda_2} (\langle n \rangle - \overline{N}_1) = \frac{1}{\lambda_2} (1 - \alpha) \left(\frac{\lambda}{\mu - \lambda} - \frac{p\alpha\lambda}{\mu^2 - p\alpha\lambda} \right). \quad (90)$$

Though (89), (90) come from an approximate solution, in the $p \rightarrow 0$ and $p \rightarrow \infty$ limits they give the correct waiting times. Taking first the limit $p \rightarrow 0$, we find

$$\lim_{p \rightarrow 0} \overline{W}_1 = \lim_{p \rightarrow 0} \overline{W}_2 = \frac{1}{\mu - \lambda}. \quad (91)$$

With $p = 0$, high and low priority customers are treated identically. The PEP reduces to a $M/M/1$ queue with arrival rate λ and service rate μ , for which the waiting time is as given by (91).

Conversely, if we make the overtake rate infinite, then high priority customers arriving at the queue will immediately overtake any waiting low priority customers. In this limit, high priority customers see an $M/M/1$ queue with arrival rate λ_1 and service rate μ , and indeed we find

$$\lim_{p \rightarrow \infty} \overline{W}_1 = \frac{1}{\mu - \lambda_1}. \quad (92)$$

The waiting time for low priority customers is

$$\lim_{p \rightarrow \infty} \overline{W}_2 = \frac{1}{(1 - \lambda/\mu)(\mu - \lambda_1)}. \quad (93)$$

This can be found by directly taking the limit, or via the requirement that $\overline{N}_1 + \overline{N}_2 = \langle n \rangle$.

Figure 6 shows \overline{W}_1 , \overline{W}_2 plotted as a function p . Again we see good agreement between simulation results and the calculated values. Increasing p interpolates between a first come first served queue ($p = 0$) and strict prioritisation according to customer class ($p \rightarrow \infty$). In designing a queueing system, one would choose p to give the desired high priority waiting time, within the constraints imposed by the asymptotic limits.

The greatest response in \overline{W}_1 , defined as the maximum value of $|\mathrm{d}\overline{W}_1/\mathrm{d}p|$, occurs at $p = 0$ in Figure 6(a), and so there is a strong relative benefit to high priority customers using even small values of p . In Figure 6(c) the value of the parameters give rise to an inflection point at $p > 0$, and hence the largest response in \overline{W}_1 occurs at some

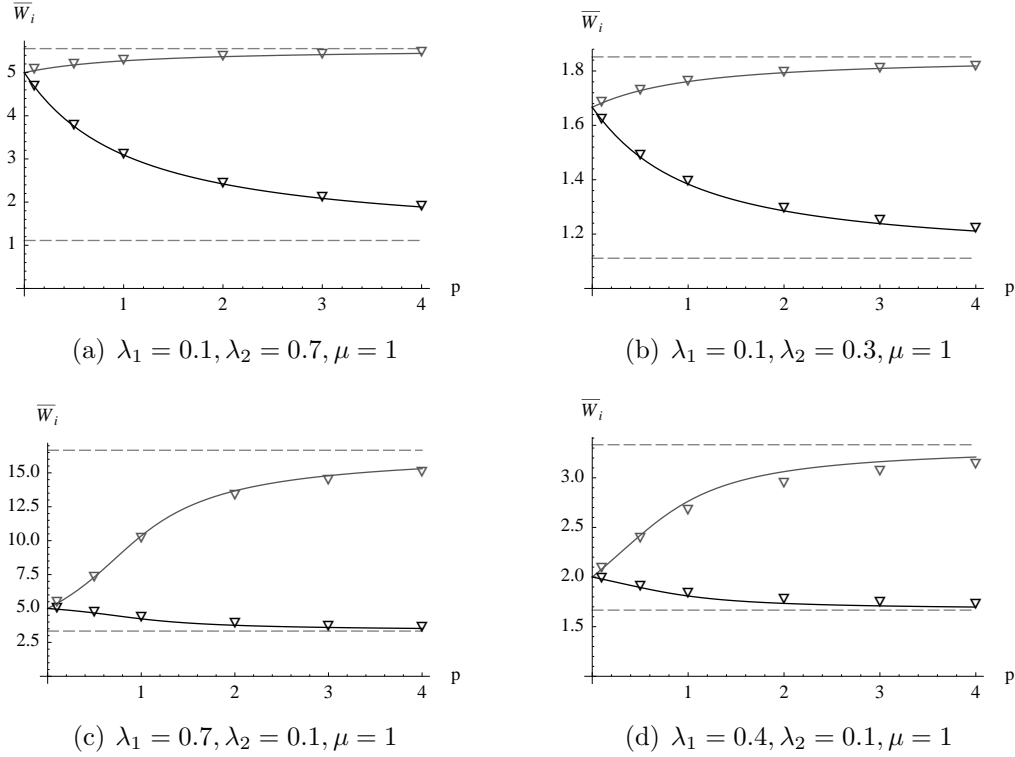


Figure 6. Average waiting time for high priority customers (black) and low priority customers (gray) plotted against overtake rate p . Dashed lines show the asymptotic values.

positive value of p . A sufficient condition for such an inflection point to occur⁺ is $d^2\bar{W}_1/dp^2|_{p=0} < 0$, which happens if and only if

$$\frac{\lambda_2}{\lambda_1} < \frac{\lambda_1 + \lambda_2}{\mu}. \quad (94)$$

Then $d^2\bar{W}_1/dp^2$ must change sign as $\lim_{p \rightarrow \infty} d^2\bar{W}_1/dp^2 > 0$. For these values of the parameters the benefit to high priority customers of switching on p is relatively small compared to the penalty for low priority customers.

4. Conclusion

In this paper we introduced a new model, the prioritising exclusion process, and analysed its stationary behaviour. In the unbounded phase, we used domain wall theory to derive the exact stationary distribution, and found a further subdivision of this phase into finite and infinite jam phases. In the bounded phase, although domain wall theory did not give exact results, it led to two complementary approximate solutions. From the direct application of the domain wall ansatz we found that the shape of the density profile could again be understood in terms of a jam, in this case either localised at the service

⁺ This is simpler than trying to solve $d\bar{W}_1^2/dp^2 = 0$.

end, or able to grow and fill the lattice. The second approach, following from a current conservation equation, allowed accurate calculation of customer waiting times.

The connection to queueing theory guided our approach to the problem and interpretation of results. The recent work on the APQ [3] points to future directions. It would be of interest to compute the complete waiting time distribution, and to compare it to that for the APQ. Additionally, the Maximum Priority Process introduced in [3] bears a strong resemblance to the jam of high priority customers in the PEP, although the exact correspondance is not clear.

Acknowledgement

We thank Peter Taylor for suggesting the PEP to us as well as for discussions, and Alexandre Lazarescu, Chikashi Arita, Guy Latouche, and Kirone Mallick for discussions and encouragement. We are grateful to the Australian Research Council for financial support.

References

- [1] Schmittmann B and Zia R K P, *Statistical mechanics of driven diffusive systems*, 1995 *Phase Transitions and Critical Phenomena* Vol. 17 (London: Academic Press).
- [2] Kleinrock L, *A delay dependent queue discipline*, 1964 *Naval Research Logistics Quarterly*, **11**(3-4) 329–341.
- [3] Stanford D A, Taylor P and Ziedins I, *Waiting time distributions in the accumulating priority queue*, 2013 *Queueing Systems* 1–34.
- [4] Spitzer F, *Interaction of markov processes*, 1970 *Advances in Mathematics*, **5**(2) 246–290.
- [5] Liggett T, 1985 *Interacting Particle Systems* (New York : Springer).
- [6] Derrida B, *An exactly soluble non-equilibrium system: The asymmetric simple exclusion process*, 1998 *Physics Reports* **301**(1–3) 65 – 83.
- [7] Schütz G M, *Exactly Solvable Models for Many-Body Systems Far from Equilibrium*, 2000 *Phase Transitions and Critical Phenomena* Vol. 19 (London: Academic Press).
- [8] Golinelli O and Mallick K, *The asymmetric simple exclusion process: an integrable model for non-equilibrium statistical mechanics*, 2006 *Journal of Physics A: Mathematical and General* **39**(41) 12679, arXiv:cond-mat/0611701.
- [9] Blythe R A and Evans M R, *Nonequilibrium steady states of matrix-product form: a solver’s guide*, 2007 *Journal of Physics A: Mathematical and Theoretical* **40**(46) R333, arXiv:0706.1678.
- [10] Derrida B, Domany E and Mukamel D, *An exact solution of a one-dimensional asymmetric exclusion model with open boundaries*, 1992 *Journal of Statistical Physics* **69**(3-4) 667–687.
- [11] Schütz G and Domany E, *Phase transitions in an exactly soluble one-dimensional exclusion process*, 1993 *Journal of Statistical Physics* **72**(1-2) 277–296, arXiv:cond-mat/9303038.
- [12] Derrida B, Evans M R, Hakim V and Pasquier V, *Exact solution of a 1D asymmetric exclusion model using a matrix formulation*, 1993 *Journal of Physics A: Mathematical and General* **26**(7) 1493.
- [13] Gwa L-H and Spohn H, *Bethe solution for the dynamical-scaling exponent of the noisy Burgers equation*, 1992 *Phys. Rev. A* **46** 844–854.
- [14] de Gier J and Essler F H L, *Bethe ansatz solution of the asymmetric exclusion process with open boundaries*, 2005 *Phys. Rev. Lett.* **95** 240601, arXiv:cond-mat/0508707.
- [15] Praehofer M and Spohn H, *Current fluctuations for the totally asymmetric simple exclusion process*, 2001 *Preprint* arXiv:cond-mat/0101200

- [16] Kolomeisky A B, Schütz G M, Kolomeisky E B and Straley J P, *Phase diagram of one-dimensional driven lattice gases with open boundaries* 1998 *Journal of Physics A: Mathematical and General* **31(33)** 6911.
- [17] Sugden K E P and Evans M R, *A dynamically extending exclusion process*, 2007 *Journal of Statistical Mechanics: Theory and Experiment* **2007(11)** P11013, arXiv:0707.4504.
- [18] Nowak S A, Fok P-W and Chou T, *Dynamic boundaries in asymmetric exclusion processes* 2007 *Phys. Rev. E* **76** 031135, arXiv:0708.0259.
- [19] Arita C *Queueing process with excluded-volume effect*, 2009 *Phys. Rev. E* **80** 051119, arXiv:0911.2528.
- [20] Cividini J, Hilhorst H J and Appert-Rolland C, *Exact domain wall theory for deterministic TASEP with parallel update*, 2013 *Preprint* arXiv:1310.2090.
- [21] Kleinrock L, 1976 *Queueing Systems, Volume II: Computer Applications* (Wiley Interscience).
- [22] Kleinrock L, 1975 *Queueing Systems, Volume I: Theory* (Wiley Interscience).
- [23] Santen L and Appert C, *The asymmetric exclusion process revisited: Fluctuations and dynamics in the domain wall picture*, 2002 *Journal of Statistical Physics* **106(1-2)** 187–199, arXiv:cond-mat/0107238.
- [24] Popkov V, Santen L, Schadschneider A and Schütz G M, *Empirical evidence for a boundary-induced nonequilibrium phase transition*, 2001 *Journal of Physics A: Mathematical and General* **34(6)** L45.
- [25] Cook L J and Zia R K P, *Feedback and fluctuations in a totally asymmetric simple exclusion process with finite resources*, 2009 *Journal of Statistical Mechanics: Theory and Experiment* **2009(02)** P02012, arXiv:0811.1543.